基于深度循环网络的声纹识别方法研究及应用 *

余玲飞 [1,2]，刘　强 [2]

(1. 浙江工商大学 杭州商学院，杭州 310018; 2. 电子科技大学 计算机科学与工程学院，成都 611731)

摘　要：声纹识别是当前热门的生物特征识别技术之一，能够通过说话人的语音识别其身份。针对声纹识别技术进行了研究，提出了一种基于卷积神经网络（CNN）和深度循环网络（RNN）的声纹识别方案 CDRNN，CDRNN 结合 CNN 和 RNN 的优势，用于移动终端声纹识别应用。CDRNN 将说话者的原始语音信息经过一系列的处理并生成一张二维语谱图，利用 CNN 长于处理图像的优势从语谱图中提取语音信号的个性特征，这些个性特征再输入到 Deep RNN 中完成声纹识别，从而确定说话者的身份。实验结果表明了 CDRNN 方案能够获得比 GMM-UBM 等其他方案更好的识别准确率。

关键词：声纹识别；深度循环网络；卷积神经网络；语谱图

中图分类号：TP391.4　　doi: 10.3969/j.issn.1001-3695.2017.07.0661

Research and application of deep recurrent neural networks based voiceprint recognition

Yu Lingfei[1, 2], Liu Qiang[2]

(1. *Hangzhou College of Commerce Zhejiang Gongshang Unversity*, *Hangzhou* 310018, *China*; 2. *School of Computer Science & Engineering*, *University of Electronic Science & Technology of China*, *Chengdu* 611731, *China*)

Abstract: Voiceprint recognition was one of the most popular biometric identification technologies, which could identify a speaker based on his voice. This paper proposed CDRNN, a voiceprint recognition scheme. CDRNN combined CNN and Deep RNN into a unified model and took advantages of both of them. For CNN was good at extracting characteristics from images, it could generate several spectrograms based on the original voice signal at first. And then, CNN would extract unique features from these spectrograms. . Finally, Deep RNN would output the speaker's identification based on these unique features. Simulation results show that CDRNN performs better than GMM-UBM and DNN-based approach.

Key Words: voiceprint recognition; deep RNN; CNN; spectrogram

0　引言

随着移动互联网的蓬勃发展和智能手机的不断普及，便捷的网络交互已成为人们日常生活中不可或缺的活动。在网络环境下，如何能准确确认交互方的身份成为日益重要的一个问题。

相比于传统的账号密码方案，基于人们自身具有的生物特征[1]的身份认证机制有着更为安全可靠的优势。人的生物特征例诸如声纹、指纹、掌纹、视网膜、人脸等，对于每个人而言具有唯一性，并且还具有稳定、不易被仿造等特点，因此得到了学术界和产业界越来越多的关注。其中声纹识别技术是根据声音对说话人进行识别，故也称为说话人识别。与指纹、视网膜等生物特征相比，声纹识别能够进行远程认证，使用成本低且易用性高。并且智能手机的全面普及，使得语音的采集也非常方便，进行认证时用户只需录制一段语音即可完成身份认证。因此声纹识别技术在金融、网络交易、国防等领域有着广泛的需求和前景[2]。

声纹识别技术的研究已有不少，早期人们对说话人识别的研究工作聚集在特征参数提取和模型匹配这两方面。从声学特征参数提取方面来看，模拟听觉特征线[25]、线性预测（linear predictive coefficients，LPC）系数、感知线性预测系数（perceptual linear predictive，PLP）[3]和梅尔频率倒谱系数（Mel frequency cepstral coefficients，MFCC）[4, 26]等参数先后被人们提出。而对于模型匹配，语音识别技术被用于人的声纹识别中。例如动态时间规整（dynamic time warping，DTW）[5]和矢量量化（vector quantization，VQ）[6]，以及人工神经网络（artificial neutral network，ANN）[7-8]等技术。

高斯混合模型（Gaussian mixture model，GMM）由于具有简单可靠和性能稳定的优点，成为声纹识别的关键方法之一[9]。基于 GMM，Reynolds 等人则提出了 GMM-UBM 模型（Gaussian mixture model-universal background model），从而将声纹识别推

基金项目：国家自然科学基金资助项目（61370204）；浙江省自然科学基金资助项目（LQ16F02001）

作者简介：余玲飞（1979-），女，浙江台州人，副教授，博士（后），主要研究方向为车载自组织网络、移动大数据（linphie@163.com）；刘强（1990-），男，硕士研究生，主要研究方向为移动大数据.

向实际应用[10]。

近年来，随着深度学习技术的发展并在图像处理、语音识别领域取得了较好的效果[11~13]，如 Palaz 等人[14]分析了卷积神经网络（convolutional neural networks，CNN）用于语音识别并取得了较好的效果。受此启发，一些研究也开始将深度学习技术应用于说话人识别[15-17]。Richardson 等人[15]将深度神经网络（deep neutral networks, DNN）用于说话人的识别，通过构建一个基于瓶颈特征（bottleneck features，BNFs）的 i-vector 系统，从语音信号中提取帧级别(frame-level)的特征。文献[16]利用 GMM 和 DNN，在具有混响的远程通话环境下，通过方言的语音特征来识别方言。文献[17]利用语音的多元音素(senone)，结合 DNN 和简化高斯概率线性鉴别分析对一段短语音信号进行建模并识别说话人。由于语音信息是一段连续的具有上下文关联的信号，而循环神经网络（recurrent neural networks，RNN）擅长对序列信号的处理，因此文献[18，19]RNN 引入对说话者进行身份识别。文献[18]利用 CTC 分类技术，对输入的语音序列进行分类并输出一段语音的 K 音素序列分布（phoneme sequence），通过对音素序列分布识别说话人。而文献[19]则进一步扩大了 RNN 的应用场景，将长短时记忆单元（long short-term memory，LSTM）引入，基于语音信号的上下文关联特征，对大规模的语音数据进行训练和识别。此外，也有一些研究工作将 CNN 和 RNN 结合来构建神经网络。如文献[22]利用 CNN-RNN 完成多标签图片分类、Fan 等人[23]将其用于基于视频的情绪感知，文献[24]则用于运动视频的事件检测等，但是用来进行声纹识别的工作几乎没有。

尽管已有不少声纹识别的研究工作，但这些工作在环境噪声、信道失配、假冒闯入、短语音等方面仍然面临着很多困难和挑战。特别是对于卷积神经网络 CNN 和循环神经网络 RNN，它们在建模能力实际上各有所长。例如 CNN 擅长图像特征提取，而 RNN 网络在时序建模上更具优势。因此，本文结合 CNN 和 RNN 的优点，提出了一种基于 CNN 和 Deep RNN 的声纹识别机制（CDRNN），同时将 CNN 和 RNN 应用于声纹识别。CDRNN 首先将说话人的原始语音转为语谱图，再利用 CNN 的结构优势从语谱图中自动提取出说话人的个性特征，随后将这些个性特征输入到 deep RNN 中完成分类，在此基础上实现实现说话人的声纹识别。

1　网络模型

人工神经网络是一个模仿生物神经网络的结构及功能的系统，由大量人工神经元组成。多个神经元排成一列从而构成神经层，多个神经层则组成人工神经网络。图 1 是一个简单人工神经网络的示意。左侧一列神经元为输入层，接收外部信号或数据；右侧一列神经元为输出层，输出系统的处理结果；两者之间为隐藏层，不为外部所观察，完成信息的处理和转换。

1.1　深度神经网络

深度神经网络 DNN 则是包含多个隐藏层的神经网络。网络模型的参数越多，表明它具有更强信息计算和存储能力，可以完成更为复杂的任务。一方面可以通过增加隐藏层的数量，从网络结构的深度方面增加网络参数；也可以通过在每个隐藏层中增加更多的神经元，从增加网络结构的宽度方面获得更多的网络参数。一般而言，增加隐藏层的数量更具优势，在增加参数的同时和能够使得网络具有更强的特征变换能力。图 2 是一个深度神经网络，该网络中含有 3 个隐藏层。

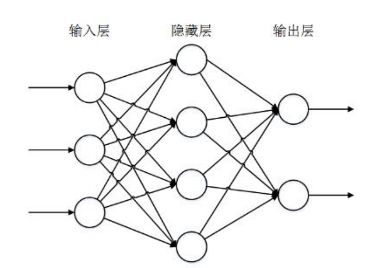


图 1　人工神经网络示意

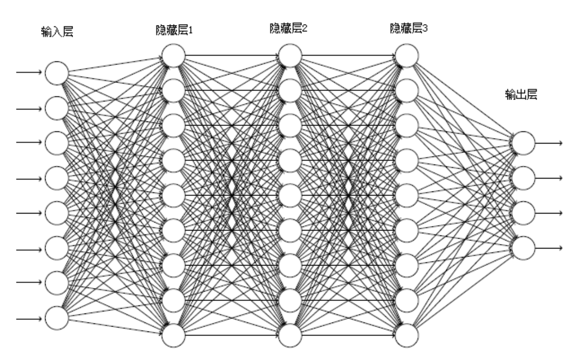


图 2　深度神经网络示意

1.2　卷积神经网络

卷积神经网络 CNN 是深度学习领域重要的网络模型之一，能在图像处理应用上取得显著的效果。CNN 是一种多层的前馈神经网络一般由若干个卷积层（convolutional layer）和池化层（pooling layer）交替构成。如图 3 所示，2 个卷积层和 2 个池化层交替构成了一个简单的卷积神经网络。

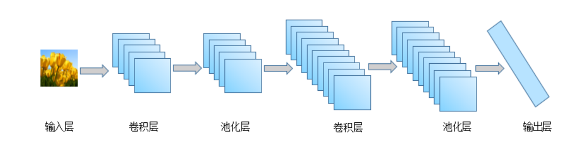


图 3　CNN 网络示意

a)卷积层。在全连接的 DNN 中，隐藏层的任何一个神经元都要和前一层所有神经元关联。但在 CNN 中，隐藏层的一个神经元仅仅与上一层中所有神经元构成的二维矩阵中的小区域进行连接。

b)池化层。卷积层用于从输入信息中提取个性特征，通常会输出维度非常高的特征，后续不便处理。此时使用池化层进行降维，简化卷积层的输出特征。同时使用池化层，输入图像具有旋转、平移和伸缩的不变特性。使用最多的是最大池化技术（max pooling），最大池化将输入图像划为多个矩形区域，分

别对每个区域提取最大值。

1.3 循环神经网络

与前馈神经网络不同，循环神经网络 RNN 则是一种反馈神经网络。RNN 的输出结果不但与当前输入信息以及网络权重有关，还与之前信息输入相关。因此，RNN 隐藏层中的神经元相互连接，同时隐藏层的输入即包括当前输入层的输出，也包括前一时刻隐藏层的输出。图 4 表示了一个简单的 RNN 网络模型

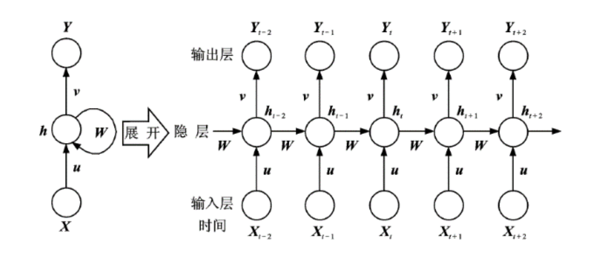


图 4　RNN 网络示意

在 t 时刻，x_t 为输入向量，h_t 为隐藏状态向量的，y_t 为输出向量，则图 4 表示的一个单隐藏层的 RNN 可定义为

$$h_t = f(W_{xh}x_t + W_{hh}h_{t-1}) \quad (1)$$

$$y_t = g(W_{hy}h_t) \quad (2)$$

其中：W_{xh} 是输入层与隐藏层之间的权重矩阵 u，W_{hh} 是隐藏层之间的权重矩阵 W，W_{hy} 则是隐藏层与输出层之间的的权重矩阵 v。通常情况下，隐藏层的激活函数有 sigmoid、tanh 和 ReLU。而输出层的激活函数一般是线性的或者是 softmax。

从理论上，RNN 能够构建长时间间隔依赖（long-term dependencies），但由于梯度爆炸等问题，仍然只能学习短周期的依赖关系。因此 LSTM 结构被引入到 RNN 中[20]。LSTM-RNN 利用 LSTM 神经元取代传统的网络神经元，即使用不同类型的门操控信息流。通过这些不同类型的门结构，LSTM 神经元可以决定何时记住输入信息，何时忘记该信息，何时输出信息。

2 CDRNN 设计

对于声纹识别应用，通常是说话人给出一段语音数据，通过对语音数据进行处理，提取出语音数据的特征（即声纹）并对其进行分类匹配，从而确定该语音数据对应的说话人的身份 ID。

2.1 声纹识别流程

图 5 显示了一个基于 CDRNN 的声纹识别系统的流程，包括三个主要的功能模块，即语音信号的预处理、语谱图的生成模块和特征提取和分类模块。其中特征提取和分类模块是整个流程中的核心模块，使用的神经网络模型结合了 CNN 和 Deep RNN 网络的优点，利用其优势互补的能力，实现说话者声纹识别的任务。

1）语音信号的预处理

由于人们发生器官的物理特性的差异，使得产生的语音信号自身的物理特性不一，而外界环境因素给语音的录制带来了噪声及其他影响，因此不能直接对原始的语音信号进行处理，必须对其进行预加重、分帧、加窗以及端点检测等信号的预处理操作。对语音数据采样量化后，首先进行预加重处理，其目的是对信号高频部分加重，减小噪声影响，使语音信号频谱平坦化；随后将一段长的语音数据划分为若干个小片段，即为分帧。这些短语音信号能够保持短时平稳状态，故可利用平稳过程方法处理；分帧带来了信号的截断效应，为了使截断处的信号能平滑过渡，需要通过加窗操作实现；最后对语音信号进行端点检测，目的是去除信号中的静音片段，保留有效的语音片段。

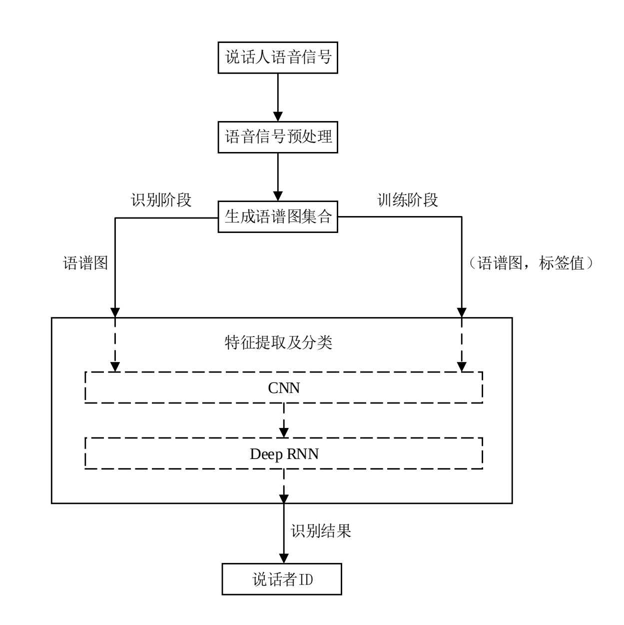


图 5　基于 CDRNN 的声纹识别流程

2）生成语谱图

语音信号的频谱实际上是随时间变化的二维图像，即语谱图，其横轴表示时间，纵轴表示信号频率成分。语谱图能够动态显示不同时刻、不同频率分量的大小，它承载的信息量远大于单时域或单频域承载的信息量。

而经过语音信号的预处理后，原始语音数据被划分为 10~30ms 长短的短语音帧。如果采用传统的滤波器来提取帧中的特征，将丢失频域上的信息，因此本文将直接生成语音信号的语谱图，保留信号的频域信息，用于后续处理。

3）特征提取及分类

特征提取是根据语谱图的信息，提取说话人声音的个性特征向量参数；而分类则是实现对该说话人语音的建模。通过一个神经网络可以统一信息的特征提取和分类，本文则利用 CNN 擅长对图像进行处理、RNN 在时序建模上具有优势的特点，将 CNN 网络和 RNN 网络统一为一个网络模型。用 CNN 网络从语谱图中提取声纹的特征参数，再通过 RNN 网络对特征信息进行时序建模。同时，具有深层结构的 RNN 网络还能够将特征参数映射到可分离空间。

2.2 语谱图生成

传统提取语音特征通常是首先对信号进行傅里叶变换，然

后使用滤波器提取特征，会导致频域信息的损失，特别是高频区域的语音信息损失更为严重。为避免频域信息的损失，CDRNN 将直接生成语音信息的语谱图。将该二维图像输入到神经网络进行处理，从中提取出语音信号的个性特征向量。

语谱图的生成过程如图 6 所示。首先得到采样量化后的语音信息，随后对语音信号进行傅里叶变换，再计算语音的能量谱密度，通过取对数和灰度图映射，将获得语音信号对应的语谱图。

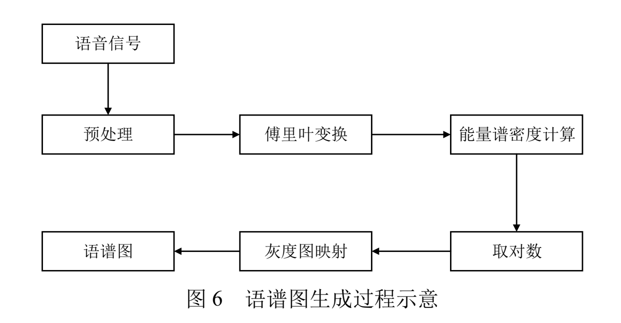


图 6　语谱图生成过程示意

由于输入至神经网络的语谱图大小固定，但不同说话者语音长度不同。因此要确保不同说话者能生成相同大小的语谱图，需要将说话者的语音信号划分为相等时长的片段，从而生成相同尺寸的语谱图。例如两个说话者 A 和 B，A 的语音信号总长度是 10 分钟，B 产生 15 分钟的语音信号。假设采样频率为 16 KHz，语音片段时长为 1s，语谱图帧长设为 512，则 A 和 B 将分别产生 600 和 900 个语音片段，并分别对应 600 张语谱图和 900 张语谱图。通过处理，每个说话者都会产生各自的语谱图，将原对语音的识别转换成对二维语谱图的识别。

2.3　网络模型设计

完成语谱图的生成后，语谱图将输入到神经网络中进行特征提取和分类，本文分别通过 CNN 和 DeepRNN 网络实现语音信号的个性特征提取及分类。

2.3.1 CNN 网络设计

如前所述，CNN 特别擅长于处理图像，而语谱图实际上就是一张二维灰度图像，图像的各种属性反映了说话者语音信号的各种特征信息。因此将语谱图作为输入，由 CNN 网络自动从输入的二维灰度语谱图中提取出语音片段的个性特征。CNN 包含多个卷积层和池化层，其中卷积层能够提取语音片段的不同特征，池化层则可以对输入的二维灰度图进行平移、缩放或其他变形操作后，仍然产生相同池化后特征，从而减少频谱变化导致的影响。

CDRNN 机制中，CNN 结构部分实际是由 n 个卷积池化单元构成，如图 7 所示，其中 n 需要根据实际情况设定。

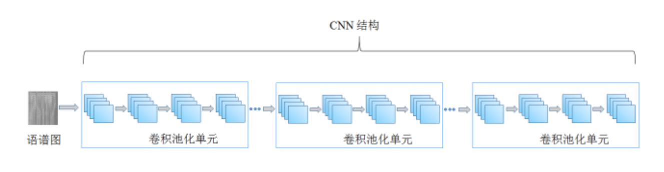


图 7　CNN 网络结构

而一个卷积池化单元实际上是一个卷积层-ReLU 层-MaxPool 层- Batch Normalization 层的结构，如图 8 所示。其中 ReLU 是激活函数，而 MaxPool 为池化函数。为了使得网络能够快速收敛，还通过 Batch Normalization 算法加速网络的训练速度。

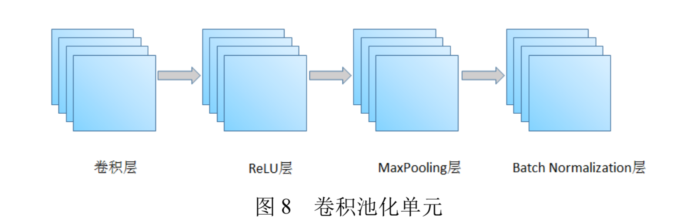


图 8　卷积池化单元

需要强调的是，对于卷积池化单元，其中的池化层在进行池化操作时，仅在频率（对应于语谱图高度）上做池化，而没有在时间（对应于语谱图的宽度）上进行池化。这主要是在时间上池化很可能导致语谱图中时序信息丢失，因此只在频率上对信号进行池化。此外，卷积池化单元和特征映射的数量、特征映射数量、卷积核大小和步长乃至池化区域大小同样需要根据具体问题和数据集通过实验进行设置。

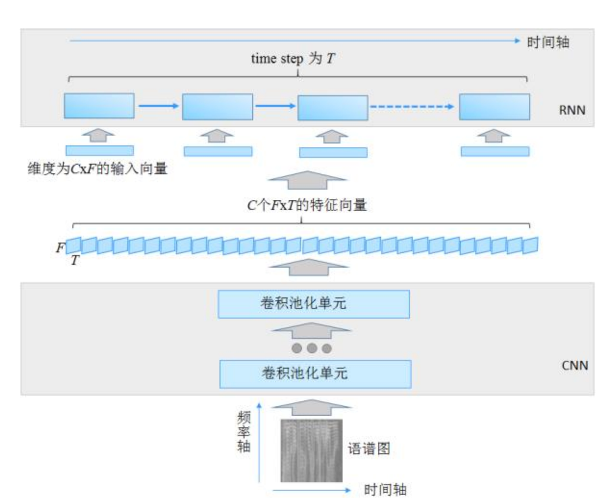


图 9　CNN 网络输出作为 RNN 网络输入

2.3.2 Deep RNN 网络设计

当 CNN 网络对语谱图的二维灰度图像处理后，其输出作为 Deep RNN 的输入完成进一步的时序建模。Deep RNN 实际上是由若干 RNN 的隐藏层进行叠加而构成，前一隐藏层的输出作为下一隐藏层的输入。相比于普通隐藏层中神经元相互独立，Deep RNN 隐藏层中包含的神经元之间则具有连接。

1)Deep RNN 输入层设计

一张二维灰度图像（语谱图）输入至 CNN 网络后，将由 n 个卷积池化单元进行处理，处理后的输出实际上是 C 张大小为 $F\times T$ 的小语谱图，其中 C 表示特征映射的数量，F 和 T 则分别是输出的小语谱图的高度和宽度。可以用一个序列来表示 CNN 网络的输出，即 $S=[S_1, S_2, \cdots S_i, S_T]$，$1\leqslant i\leqslant T$，而序列中的元素 S_i 则是一个大小为 $C\times F$ 的向量。也就是说 CNN 将输出 T 个大小为 $C\times F$ 的向量，这些向量作为作为 RNN 网络的输入，它们之间有一个对应关系，即 CNN 网络输出序列 S_i 作为 RNN 在 i

时刻的输入。也就是说，RNN 在 i 时刻的输入是一个 $C\times F$ 维的向量，它的步长则等于 T。图 9 显示了 CNN 的输出序列和 RNN 输入之间的对应关系。

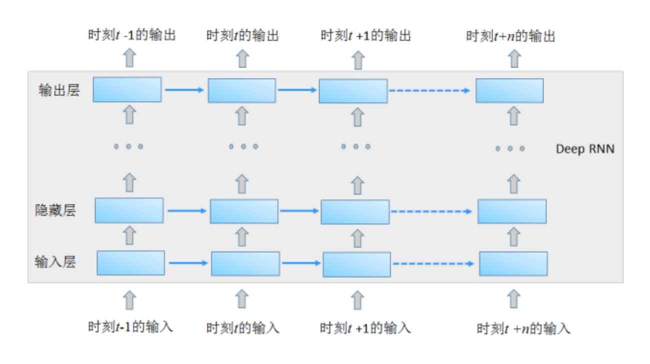


图 10　Deep RNN 网络结构示意

2）RNN 隐藏层设计

Deep RNN 是由多个 RNN 堆叠起来的，其中每一层的输出序列作为下一层的输入序列，其结构如图 10 所示。和传统的神经网络相比，Deep RNN 的特点是在每一层都会有时间上的反馈循环。对于实际问题，Deep RNN 中的隐藏层通常使用改进的 RNN 如 LSTM-RNN 或 GRU-RNN，它们解决了基本 RNN 中缺乏长期依赖关系的问题，这可以使神经网络能够记住更长一段时间跨度的输入数据。对于 Deep RNN 网络，设计隐藏层的结构时要考虑两个参数，即隐藏层层数的多少和隐藏层中神经元的节点数量。这两个参数同样需要依据实际需求进行设定。一般而言，在相同参数个数的倾向下下，设置更多的层数比增加每层更多的节点数能够获得更好的效果。

3）Deep RNN 输出层

Deep RNN 的输出层比较简单，就是使用一个 softmax 分类器进行分类，通过 softmax 分类，使得输出层的节点数对应于说话人的人数。

2.3.3 网络模型训练

CDRNN 模型的训练采用了监督学习的方法，首先要对所有的数据打上标签，然后把数据和所对应的标签作为训练集。假设待训练的语音信号为 K 个（由 K 个说话人产生），第 i 个语音信号生成的语谱图序列为 S_i=（$S_i^1, S_i, \cdots S_i^m$），m 为该语音信号生成的语谱图数量，则第 j 张语谱图 S_i^j 对应着一个二维矩阵，给其一个标签值为 i-1，这意味着同一说话人的所有语谱图具有相同的标签，而此标签则可标识该说话人的 ID，S_i^j 和它的标签 i-1 则构成一个训练样本（S_i^j，i-1）。

训练样本进行训练前，还需对样本数据进行标准化或归一化处理，即将数据按一定比例缩放，将数据映射为一个小区间内，从而去除数据的单位限制，将数据转换为无量纲数值。同时，数据标准化后还能够提高模型收敛速度和准确度。本文采用机器学习中常用的 Min-Max 标准化机制对二维灰度图像的每个像素进行标准化，数据标准化后，像素点取值区间为[0,1]。

经过给样本数据打标签和数据标准化后，则可开始对样本数据进行训练。对多个样本语音信号训练的过程实际上是一个多分类的任务。CDRNN 选择的代价函数是交叉熵函数，同时利用了 BP 及 BPTT 算法计算梯度，从而完成样本数据的训练。

2.3.4 网络模型识别

网络模型对语音数据集进行训练，训练完成了即可用于声纹的识别。进行识别时，首先说话人产生一段测试语音信号，该信号经过预处理后生成了 N 张语谱图，这些语谱图同样要进行数据标准化，然后再依次将数据标准化后的语谱图输入到 CDRNN 网络模型中，模型最终回给出每一张语谱图所对应的说话者的身份 ID。显然，N 张语谱图会输入 N 个说话者的 ID，而测试语音对应的声纹所属的说话人 ID 则被认为是这 N 个 ID 中出现次数最多的那个 ID。

3　仿真实验

3.1　实验设置

实验平台采用了 Google 的开源深度学习框架 TensorFlow[21]，在 TensorFlow 平台上对样本数据进行训练，训练好的模型可以移植到移动手机上，移动手机则可对说话者进行语音采样并通过训练好的模型进行声纹的识别。对样本数据进行训练的数据为 DELLC4130 服务器，配置了 4 块英伟达 Tesla GPU，显存大小为 24G。

3.1.1 语音数据集

实验所使用的语音数据是从真实环境中进行采集的。通过智能手机对 40 个不同的学生各自录制了 10~20 分钟的语音数据。由于环境因素的影响，采集的语音信号中不可避免的包含了背景噪声数据。每个学生的语音数据被划分为 1s 时长的语音片段，这些语音片段的前 80%的数据作为训练数据集用于网络模型训练，而后 20%的数据则作为测试数据集对训练后的网络模型进行测实验证。此外，定义识别率作为性能评价指标，即识别正确的语音片段的数量与测试数据集中语音片段的总数量的比值。

3.1.2 语谱图参数

对每个语音片段生成语谱图时，帧长设为 512，那么生成语谱图后将得到 256 个像素点，这对应语谱图高度。实际上实验时仅取了前面的 128 个像素点，因为语音信号频率一般在 300-3000Hz 区间，在区间外的信号是噪声信号，可以忽略。而另一参数帧移设置为 160，由于采样频率是 16KHz，则 1s 时长的语音片段将产生 16K 个采样点，故能得到 100 帧，意味着语谱图宽度为 100 个像素点。因此最终生成的语谱图大小为 128×100 ，即高度是 128 个像素点，宽度为 100 个像素点。

3.1.3 CNN 结构参数

CNN 的参数如卷积池化单元数量、步长、卷积核大小和特征映射数量等需依据实际数据集的调参来确定。经实际调参，CNN 结构的参数设置如下：

a）卷积池化单元的数量 n=4，第一个池化单元的特征映射数量设为 32，而后三个池化单元中特征映射的数量则设置为 64。

b）卷积层中卷积核大小为 5×5，步长设为 1×1，并同时

在频率方向上和时间方向上均进行卷积操作。

c）池化层中，池化区域的大小设为1×1，步长仍为1×1，仅在频率方向上进行池化。

3.1.4 Deep RNN 结构参数

Deep RNN 的两个重要参数即为 RNN 的层数以及每层的节点数。RNN 层数越多，识别说话人 ID 的能力就越强，但层数多意味着训练开销大，并较易产生过拟合现象。通过对这两个参数的不同组合获得不同的 RNN 结构，并测试不同网络结构下的识别率，选择识别率最高的网络结构对应的 RNN 层数和每层节点数作为 Deep RNN 参数。

如图 11 所示，RNN 的层数分别为 1，3，5 和 7，每层的节点数则为 128，256 和 512，这样共可获得 12 种组合，对应 12 个网络模型。由图可见，随着 RNN 层数的增加，系统的识别率基本呈上升趋势。类似地，当 RNN 层数不超过 5 时，每层的节点数越多，识别率也就越高。但是 RNN 层数为 7，每层节点数为 512 时，其识别率反而低于每层节点数为 256 时的识别率。这说明并非层数和每层节点数越多，识别结果就越好。其原因在于随着层数和每层节点数的增加，参数数量几何级数上升，而训练集大小有限，就容易导致过拟合现象。基于实验结果，将 RNN 层数设置为 7，而每层节点数设置为 256。

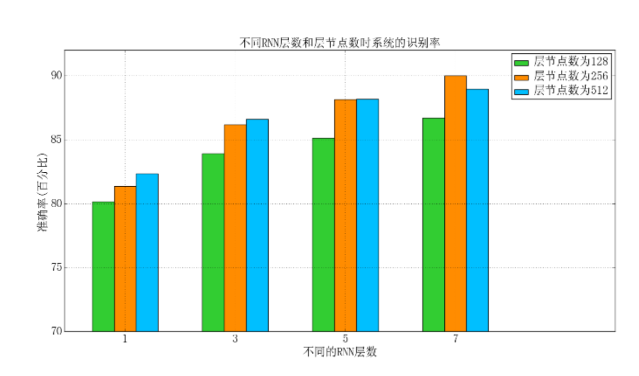


图 11　Deep RNN 中不同网络参数的识别率

3.2　实验结果

首先比较了基于 CDRNN、GMM-UBM[10]、DNN[17]和 GMM-DNN[16]的四种机制在本语音数据集上的识别准确率，结果如图 12 所示。显然，随着说话者人数的增加，四种机制的识别准去率均有所下降，而 GMM-UBM 的识别率下降非常快，这是由于 GMM-UBM 中关键参数混合度的取值对结果又较大影响。而 CDRNN 的识别率则下降较慢，且比 GMM-UBM 的识别率高约 18%左右，特别是在说话人数量较多的时候。此外，CDRNN 也比 DNN 和 GMM-DNN 高约 6%，说明后端使用 RNN 后，能够获得比使用 DNN 更好的结果。

本质上，CDRNN 使用的是 CNN+RNN 这样的前后端网络模型，前端是 CNN，后端是 RNN。将 CNN+RNN 的网络模型和仅使用 RNN 建模以及前端采用 DNN、后端使用 RNN 的深度网络模型进行了性能比较。

图 13 表示了上述三个模型在 RNN 层数为 1、3、5,每层节点数为 256 个时的识别准确率。可以看出，随着 RNN 层数的增加，三种网络模型的识别准确率都得到了一定程度的提高，而只使用 RNN 网络模型的识别准确率最低。在后端使用相同的 RNN 网络的前提下，前端采用 CNN 网络获得的识别率要比前端采用 DNN 网络的识别率更高一些。

随后将网络模型的前端固定为 CNN 网络，后端则分别为 DNN 和 RNN 网络，其层数分别为 1、3、5，每层的节点数分别为 128、256 和 512，得到的结果如图 14 所示。

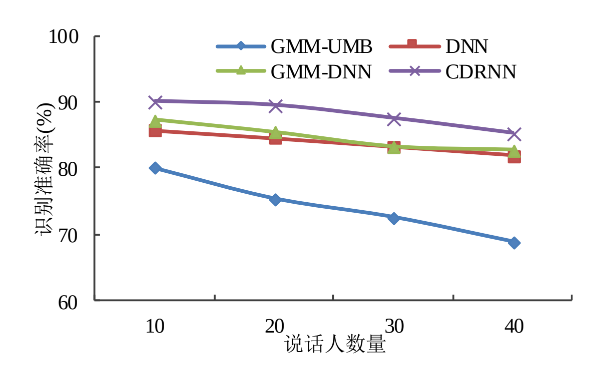


图 12　四种方案的性能比较

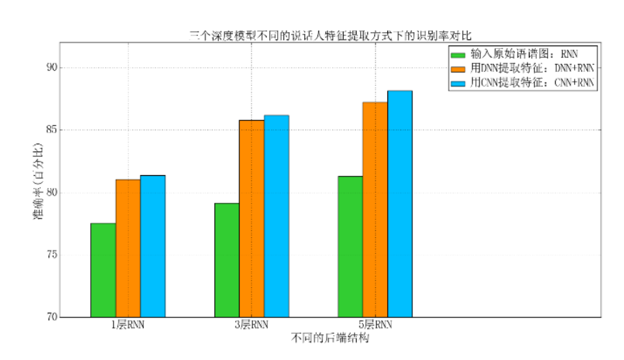


图 13　RNN、DNN+RNN 和 CNN+RNN 模型的识别率

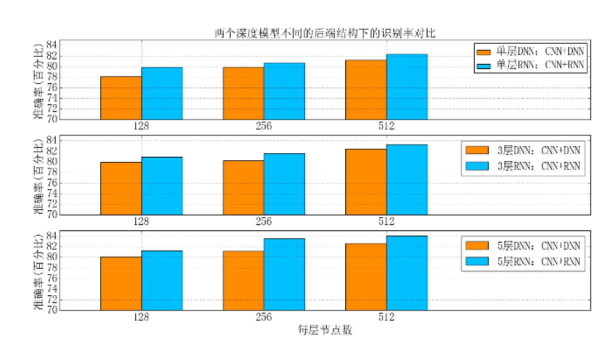


图 14　CNN+DNN 和 CNN+RNN 模型的识别率

由图可见，识别率同样随着层数和每层节点数的增加而增加。而无论两个模型的后端网络的层数、每层的节点数如何变化，当参数相同时，CNN+RNN 网络模型获得的识别率要笔 CNN+DNN 高约 4%左右，体现了 CNN+RNN 结构的优势。

4　结束语

本文利用 CNN 处理图像能力强以及 RNN 网络易于对时序数据进行建模的特点，提出了 CDRNN 机制，结合了 CNN 和

RNN 的优势，将其用于声纹识别应用。通过真实语音数据集，利用 CDRNN 进行训练和测试，对声纹识别的准确率高于其他方案。

参考文献：

[1] Jain A, Ross A, Prabhakar S. An introduction to biometric recognition [J]. IEEE Trans on Circuits & Systems for Video Technology, 2004, 14 (1): 4-20.

[2] Furui S. Recent advances in speaker recognition [J]. Pattern Recognition Letters, 1997, 18 (9): 859-872.

[3] Hermansky H. Perceptual linear predictive (PLP) analysis of speech [J]. Journal of the Acoustical Society of America, 1990, 87 (4): 1738-52.

[4] Vergin R, O'Shaughnessy D, Farhat A. Generalized Mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition [J]. IEEE Trans on Speech & Audio Processing, 1999, 7 (5): 525-532.

[5] Dutta T. Dynamic time warping based approach to text-dependent speaker identification using spectrograms [C] : Proc of the Congress on Image and Signal Processing, 2008. New York, USA, 2008: 354-360.

[6] Gray R. Vector Quantization [J]. IEEE ASSP Magazine, 1990, 1 (2): 75-100.

[7] Gardner M W, Dorling S. Artificial neural networks-a review of applications in the atmospheric sciences [J]. Atmospheric Environment, 1998, 32 (14–15): 2627-2636.

[8] Jain A, Mao J, Mohiuddin K. Artificial neural networks: a tutorial [J]. Computer, 1996, 29 (3): 31-44.

[9] Reynolds D, Rose R. Robust text-independent speaker identification using Gaussian mixture speaker models [J]. IEEE Trans on Speech & Audio Processing, 1995, 3 (1): 72-83.

[10] Reynolds D, Quatieri T, Dunn R. Speaker Verification Using Adapted Gaussian Mixture Models [J]. Digital Signal Processing, 2000, 10 (1-3): 19-41.

[11] Schmidhuber J. Deep learning in neural networks: an overview [J]. Neural Networks, 2014, 61 (3): 85-94.

[12] Abdel-Hamid O, Mohamed A, Jiang H, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2012: 4277-4280.

[13] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition [J]. Computer Science, 2014, 13 (2): 120-131.

[14] Palaz D, Magimai M, Collobert R. Analysis of CNN-based speech recognition system using raw speech as input [C]// Proc of International Speech. 2015: 11-15.

[15] Richardson F, Reynolds D, Dehak N. Deep neural network approaches to speaker and language recognition [J]. IEEE Signal Processing Letters, 2015, 22 (10): 1671-1675.

[16] Phapatanaburi K, Wang L, Sakagami R. Distant-talking accent recognition by combining GMM and DNN [J]. Multimedia Tools & Applications, 2016, 75 (9): 5109-5124.

[17] Kanagasundaram A, Dean D, Sridharan S, Fookes C. DNN based Speaker Recognition on Short Utterances [C]// Proc of Speaker & Language Recognition Workshop. 2016

[18] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. 2013: 6645-6649.

[19] Sak H, Senior A, Beaufays F. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition [J]. Computer Science, 2014, 13 (8): 338-342.

[20] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9 (8): 1735-1780

[21] TensorFlow [CP/OL]. http://www. tensorflow. org.

[22] Wang Jiang, Yang Yi, Mao Junhua, et al. CNN-RNN: a unified framework for multi-label image classification [C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2285-2294

[23] Fan Yin, Lu Xiangju, Li Dian, et al. Video-based emotion recognition using CNN-RNN and C3D hybrid networks [C]// Proc of the 18th ACM International Conference on Multimodal Interaction. 2016: 445-450

[24] Jiang Haohao, Lu Yao, Xue Jing. Automatic soccer video event detection based on a deep neural network combined CNN and RNN [C]// Proc of the 28th IEEE International Conference on Tools with Artificial Intelligence. 2016: 490-494

[25] 林琳，陈虹，陈建. 基于鲁棒听觉特征的说话人识别 [J]. 电子学报，2013, 41 (3): 619-625.

[26] 曹洁，余丽珍. 基于 MFCC 和运动强度聚类初始化的多说话人识别 [J]. 计算机应用研究, 2012, 29 (9): 3295-3298.